UCRL-ID-132850

An LLNL Perspective on ASCI Data Mining and Pattern Recognition Requirements

C. Baldwin
C. Kamath
R. Musick

January 1, 1999

Lawrence Livermore National Laboratory

Work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract W-7405-ENG-48.

An LLNL Perspective on ASCI Data Mining and Pattern Recognition Requirements

A Working Document

Chuck Baldwin, Chandrika Kamath*, Ron Musick
{baldwin5, kamath2, musick2}@llnl.gov

January 1999

Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
UCRL:

This working document has been put together by the members of the Sapphire project at LLNL. The goal of Sapphire is to apply and extend techniques from data mining and pattern recognition in order to detect automatically the areas of interest in very large data sets. The intent is to help scientists address the problem of data overload by providing them effective and efficient ways of exploring and analyzing massive data sets. One of the key areas where we expect this technology to be used is in the analysis of the output from ASCI simulations. It is expected that a simulation running on the 100Tflop ASCI machine in the year 2004 will produce data at the rate of 12TB/hour. Given the difficulties we currently have in analyzing and visualizing a terabyte of data, it is imperative that we start planning now for ways that will make the analysis of petabyte data sets feasible.

This document focuses on the relevance of data mining and pattern recognition to ASCI, discusses potential applications of these techniques in ASCI, and identifies research issues that arise as we apply the algorithms in these areas to massive data sets.

1. Introduction

There are several different facets to the task of scientific data exploration, including high performance storage and networking, data management, and data visualization. In addition to providing hardware support for storage, ASCI is currently addressing the problems faced by scientists who are working with inadequate visualization and data management tools. Efforts in these areas include a scalable visualization architecture to enable interactive visualization of massive data sets, and an environment that provides intelligent assistance in managing (searching, browsing, organizing, and annotating) terascale scientific data sets.

While these efforts go a long way towards addressing the problem of data overload, they still leave the scientist looking at large amounts of data manually, frame by frame. Visualization allows the scientist to look at only a limited number of dimensions at a time, which may become a problem if the data is multi-dimensional. In addition, visual data analysis is inherently subjective, and there are human limitations in absorbing details when presented with a large amount of visual information. These constraints make it imperative that we use automated techniques, such as those from data mining and pattern recognition, to assist in the process of data exploration.

---

* Author to whom correspondence must be addressed.

Data mining (Figure 1) is an iterative and interactive process that automatically identifies patterns, associations, and other statistically significant structures in data. In addition to enhancing our understanding of known phenomena, it can also help us in knowledge discovery, that is the identification of new phenomena. Pattern recognition, a key step in data mining, is the discovery and characterization of patterns, which are defined as arrangements or orderings with an underlying structure. Patterns in data are identified using features, which are measurable attributes that have been extracted from the data.

In the last few years, data mining has been extensively used in commercial applications, such as market basket analysis, directed marketing, and credit card fraud, to extract useful information from large volumes of data. This work has not gone unnoticed by scientists who are faced with similar problems of data overload. Recent successes in projects such as SKICAT and JARtool have demonstrated that these algorithms can be used equally well in the analysis of scientific data. Therefore, data mining and pattern recognition techniques hold great promise as tools that will enable the automated exploration of petabytes of ASCI data.

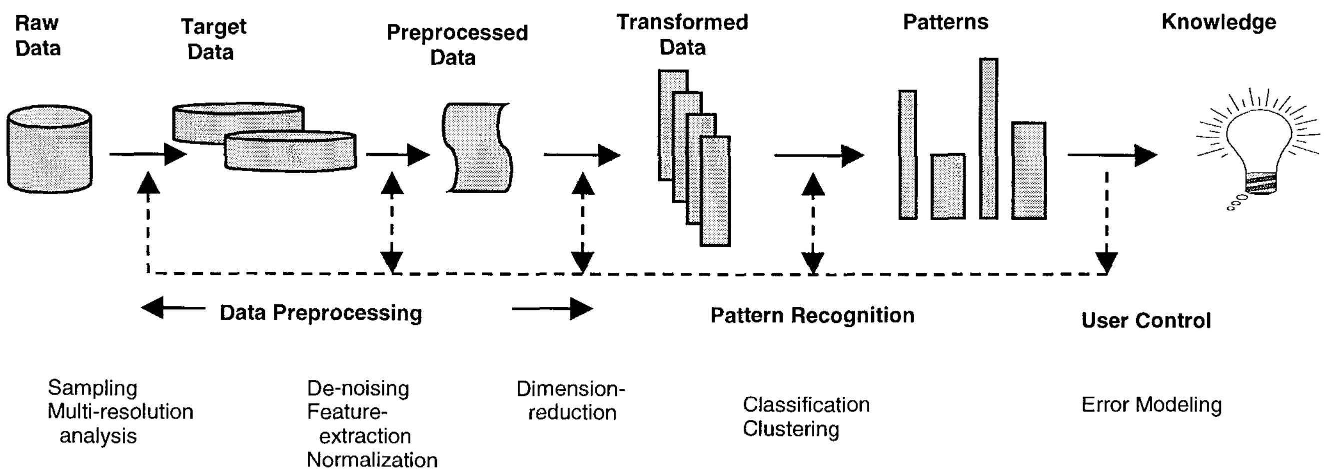

Figure 1: The data mining process

The figure above illustrates the overall process of data mining; it also includes the operations that may be performed during each transformation step. The output from any one step could indicate possible refinements in one of the previous steps, making the process iterative. Further, as some of the steps are domain and problem dependent, it is important that the domain scientist be closely involved with each step of the entire data mining process.

Starting with raw data, we first preprocess it to extract key features. This multi-stage pre-processing step is a time-consuming, but critical, first step in data mining and crucial to the success of the process. It simplifies the pattern recognition problem without throwing away important information. Some of the operations performed in this step are domain and problem dependent. For example, if we are comparing experimental images with computational images, image-processing techniques will play an important role in data pre-processing. Or, depending on the pattern of interest to the scientist, we may need to extract different features from the output of a simulation.

The first step in data pre-processing converts the raw data into target data on which data mining techniques are applied. This step allows us to work with smaller data sets through the use of sampling and multi-resolution analysis using wavelets. The next step extracts features from the data using domain-dependent techniques such as de-noising, feature extraction, and feature normalization. Next, we use various dimension reduction techniques to extract key features in order to make the problem tractable for multi-dimensional data sets. In each of these steps, there are several options available depending on the data, such as the use of different types of wavelets for de-noising image data, or various linear and non-linear techniques for dimension reduction. The key challenges in the data pre-processing steps include the identification of techniques that can be applied to scientific data in the form of images or mesh data, the generation of scalable algorithms that can be applied to large-scale data, and the efficient parallel implementations of these algorithms.

Once the data has been pre-processed, we can use pattern recognition algorithms to identify the patterns of interest. We broadly define two categories of pattern recognition: *classification* or the learning of a function that maps a data item into one of several pre-defined classes, and *clustering*, which is a descriptive task where one seeks to identify a finite set of categories or clusters to describe the data. The problems posed by pattern recognition algorithms for tera- and peta-byte data sets are particularly challenging. For example, is it possible to modify existing algorithms, or design new ones, that are scalable, robust, accurate, and interpretable? Further, can these algorithms be applied effectively and efficiently to complex, multi-dimensional data? And, is it possible to implement these algorithms efficiently on large-scale multi-processor systems so that a scientist can interactively explore and analyze the data?

In the final step of data mining, the patterns are presented to the scientist in a way that is easy to interpret and understand. In order to interactively mine massive data sets, the user may be interested in controlling the process by exploiting possible trade-offs between accuracy and computational effort. Using techniques from error modeling we would like to provide scientists with the ability to control the accuracy of the results, as well as a measure of confidence in the results obtained through data mining.

2. Description of output generated by ASCI simulations

Most of the large data sources in ASCI are data generated from computational codes or simulations. This data is primarily mesh data; it is generated when continuous variables are approximated by their values at discrete points on a mesh that overlays the region where the phenomena of interest is being simulated. A mesh provides a way of breaking a surface or volume down into an interconnected grid of much smaller zones. Each zone stores a variety of computed or collected variables, some of which constitute the physical quantities being modeled and others diagnostic or informational data. The expectation is that if the zones are small, the well-understood micro-scale properties and interactions can be modeled with enough accuracy to provide precise predictions of macro-scale events. Two important issues that effect the amount of data generated by these simulations are the fidelity, or resolution, of the simulation in both time and space and the number of simulations required to gain a level of confidence that the simulation does indeed model the phenomena being simulated accurately. The first issue has both numerical and qualitative considerations. Accuracy and stability requirements for the various numerical methods used, along with the qualitative judgement that the resolution is sufficient to provide significant information, lead to simulations with very large zone count and time step requirements. The second issue is a realization that the simulations are approximations to the true physical phenomena being modeled. To gain a level of confidence in the output from the computational codes, many simulations are performed, each with various input data perturbed in a relevant fashion. Information gained by analyzing this output can be used to either make decisions or define further parameter modifications.

Mesh data is typically stored in highly structured binary flat-files, using standard self-describing formats like NetCDF or HDF. The mesh can be accessed through high-level APIs (such as LLNL's Silo or Sandia's Exodus) that provide methods to read and write individual components of the mesh directly, without reading the rest of the structure into memory. The data itself is also highly structured. The 2 and 3-dimensional zones in a mesh are made up of lists of points, or nodes. For example, a cubic zone has 8 nodes that describe the corners of the cube. These nodes have X, Y, and Z coordinates, whose values can change if the particular mesh deforms over time. Variables can be assigned to the zone centers, faces, edges, or nodes. Variable values can be recorded at each time step in a simulation and for each zone in the mesh.

3. Potential applications of data mining in ASCI

There are two main areas where we expect data mining and pattern recognition to contribute to the ASCI vision:

3.1 Objective comparison of computational to experimental data

ASCI makes extensive use of computational models to understand and predict the behavior of complex physical phenomena. The verification and validation of these models involves comparing the output from a series of simulations with each other, or with experimental data in the form of images. These experimental images, however, can be noisy, few in number, and difficult to characterize precisely. Currently, comparisons that determine the strength of the fit between experimental and computational data are made based on the experience, or the "gut feeling" of scientists. Needless to say, it can be difficult or impossible for the scientist to quantify gut feeling in terms of precise mathematical algorithms.

Pattern recognition and stochastic modeling techniques can be used to construct objective figures of merit that supplement or replace current subjective metrics. Computable figures of merit will improve the quality and repeatability of comparisons, and may help speed overall model evaluation by automatically filtering out large sets of computational data that are not close enough to the experimental data to warrant the scientist's attention.

3.2 Identification of regions of interest in computational data

It is expected that the output from ASCI simulations in the year 2004 will be measured at the rate of 12TB/hour. In order to analyze and understand these massive data sets, scientists will need to interact with the data in novel ways. Conventional ways that have worked with megabytes of data may no longer be applicable to petabyte-size data sets.

The ASCI visualization effort is already addressing part of this problem through the use of a prepare, browse, and select strategy that will enable interactive visualization of the data. However, as explained in the introduction, visualization by itself may not provide a complete solution to the problem of interactive exploration of massive data sets. We can complement tera-scale visualization by providing an automated aspect to data exploration through the use of data mining and pattern recognition techniques. These methods can be applied in two ways:

- Identification of known but poorly defined patterns: In this case, there is no mathematical formulation that describes the pattern or the region of interest. It is therefore difficult to detect them by using a code segment in the simulation code. However, scientists can easily identify these patterns, such as fluid flow instabilities, when they look at the output from a simulation. Examples of such areas of interest can be used to train classification algorithms to automatically detect other similar regions of instability
- Computer-assisted serendipity: Here, the interest is in previously unknown or undefined patterns, such as possible anomalies in the output. These can be identified either as outliers by a clustering algorithm, or as a cluster of data items that are associated together for an

unknown reason. Further exploration of these data items could potentially enhance our understanding of the phenomena being simulated.

Having identified regions of interest in large data sets using pattern recognition, we can then use traditional visualization and data exploration methods in these smaller regions.

4. Research issues

This section outlines some of the research issues of note in addressing the tasks introduced above. This list is not comprehensive, and will be expanded as additional information is gathered.

4.1 Storage and access to large datasets

ASCI datasets will be measured in the terabytes, or even petabytes. One must deal with data storage, and transport across networks, through the storage hierarchy, from tertiary storage to disk to main memory. These issues dealing with the physical transfer of data are being addressed by the data storage effort within ASCI.

4.2 Mining massive data sets

In order to implement pattern recognition techniques on petabyte data sets, innovative ways will be needed to make the problem tractable. Sampling the data is one option; however, if the pattern of interest belongs to a class with low probability, there is the possibility that it may be completely overlooked. And, in the case of knowledge discovery, the way in which the data is sampled could very well influence what we find in the data. Another solution to the problem of massive data sets is the use of multi-resolution techniques, which then raises the question of feature extraction from coarser-resolution data in order to identify patterns in the finer-resolution data. The Computational Thumbnails project at LANL (See Section 5) is of relevance here.

4.3 Feature extraction from complex data

The extraction of relevant features is a key step in data mining. It often determines the success or failure of pattern recognition methods, and caution must be exercised while processing the data to extract features. There are several constraints, outlined below, that must be addressed for feature extraction to be successful.

Scientific applications generally produce data of various types: byte (interpreted as gray scale images), integer (electronic counters or CCD output) or floating point (computer simulation data). This data usually has a physical meaning associated with it and transformations that remove or even obscure this meaning should not be applied. In addition, the application and the problem being addressed may dictate which data transformations are acceptable and which are not. Therefore, care must be exercised during data pre-processing to ensure that no useful information is lost.

In ASCI, we expect to deal with both data from sensors and output from computations. Data collected from sensors can be noisy, with built-in skew that reflects the physical properties of the sensor, rather than the actual characteristics of the phenomena being measured. Similarly, computational output can be skewed from assumptions built into the computer codes reflecting our incomplete understanding of the domain, or tradeoffs for computational efficiency. The effects of this noise and skew need to be understood and eliminated wherever possible. This can be accomplished by computational techniques such as wavelets to reduce the noise and normalization to remove possible skew caused by the artificial units used to measure the data.

Pattern recognition algorithms typically operate on data consisting of rows of feature vectors, one for each data item. Before they are transformed into feature vectors, the native data models used in ASCI often result in highly structured, 2D or 3D objects with complex, inter-related components and inherited properties. Mapping from native data into feature vectors can explode the number of explicit features for each data item. It can potentially lose information as well.

In addition, derived data is frequently calculated from the output of a simulation during visualization. Can data mining use this derived data in order to enhance the feature set?

4.4 High dimensional data

Often, physical phenomena are described using a large numbers of features. This problem of high dimensionality is usually solved using dimension reduction techniques such as principal component analysis, projection pursuit etc. Unfortunately, these methods are linear, and may therefore be limited in their applicability. We will need to investigate non-linear techniques such as Kohonen self-organizing maps to see if they make it easier to identify key features in multidimensional data.

4.5 Efficient and effective pattern recognition algorithms

One of the open research problems in large-scale data mining is the scalability of classification and clustering algorithms to large, complex, multi-dimensional data sets. Traditional clustering algorithms tend to fail if the number of features is very large. In the case of classification, the complexity of an algorithm may be very high. In many cases, such complexity is not warranted. We need further investigation into more efficient algorithms, as well as scalable, parallel, out-of-core implementations of these algorithms.

4.6 Error modeling and the tradeoff between accuracy and computational time

Identifying the sources and magnitudes of error in data and understanding how they interact to affect interpretations drawn from the data is a critical component of any scientific endeavor. Error modeling can be used to improve the interpretability of results as well as the quality of decisions made from the data. It can also provide a measure of the confidence one can have in the end result.

Understanding errors can also be useful in the interactive exploration of large data sets. This task is computationally intensive, and to facilitate interaction, we may need to give the user the ability to control the tradeoff between accuracy and computational time. This would allow a user to request results faster, at the cost of a (specified) lower degree of confidence in the result. Predicting the computational cost required to achieve a user-specified expected degree of accuracy, or likewise quantifying the accuracy achieved after computing for a fixed amount of time is an open area of research.

4.7 Scalable parallel implementations

Given the amount of data that will be generated as a result of ASCI, fast and robust techniques will be necessary in order to meet the time constraints expected for analyzing the data. Techniques that are parallel, or capable of being made parallel, will be significant in all phases of the data mining process. Techniques which are multi-scale or hierarchical in approach such as wavelets, wavelet packets, and pyramidal algorithms have proven very useful in data analysis and could provide the means of making the data mining process scalable. This idea of focusing on scalable techniques will help us attain our goal of mining petabyte data sets in 2004, while working with terabyte size data in the interim.

4.8 Leveraging other relevant efforts

Any effort to use data mining techniques in ASCI must leverage existing efforts that are relevant such as the ASCI scientific data management project, the work on data models and formats, the visualization architecture etc. Some of these may introduce additional research issues. For example, can we use the data compression work that is already being done under ASCI, and if so, how will it effect the extraction of features from the data?

5. Relevant projects, and contact points

This is an incomplete list of other relevant tri-lab projects in the area of data exploration, along with the name of the PI, and the URL for the web page describing the project. This list will be updated whenever we are provided new pointers to projects at LLNL, SNL, and LANL. Web links that are in *italics* are under construction, and will be made available shortly

Sapphire

Large-scale data mining and pattern recognition
Chandrika Kamath (kamath2@llnl.gov)
http://www.llnl.gov/CASC/sapphire/

SAVAnTS

Scalable algorithms for the visualization and analysis of tera-scale data
Mark Duchaineau (duchaineau1@llnl.gov)
http://www.llnl.gov/CASC/savants/

IA:

Intelligent Archive
Jeff Long (long6@llnl.gov)
http://www.llnl.gov/ia/

Avatar

Adaptive visualization aid for touring and recovery
Philip Kegelmeyer (wpk@ca.sandia.gov)
http://www.ca.sandia.gov/avatar/

Computational thumbnails

Generation and use of compressed representation of tera-scale data sets
David Sigeti (sigeti@lanl.gov)
http://www.lanl.gov/projects/sdm/

NWIG

Metadata search and organization efforts
Bruce Lownsbery (lownsbery1@llnl.gov)
http://www.llnl.gov/nwig/ (Restricted access)

DataFoundry

Scientific data management for the Human Genome project
Ron Musick (rmusick@llnl.gov)
http://www.llnl.gov/CASC/datafoundry/

MeshTv

Visualization tool for 2- and 3-dimensional meshes
Eric Brugger (brugger1@llnl.gov)
http://www.llnl.gov/bdiv/meshtv/

ASCI data formats

Data models and formats for ASCI data
Mark Miller (miller86@llnl.gov), Larry Schoof (laschoo@sandia.gov)
John Ambrosiano (ambro@lanl.gov)
http://www.ca.sandia.gov/asci-sdm/

ASCI SDM

Scientific data management for ASCI
Celeste Matarazzo (celestem@llnl.gov), Philip Kegelmeyer (wpk@ca.sandia.gov)
John Ambrosiano (ambro@lanl.gov)
http://www.ca.sandia.gov/asci-sdm/

6. Acknowledgments

This document has benefited from many discussions with those listed in the above projects.

The web page already carries the LLNL auspices statement.